Environmental Health Perspectives
Vol. 87, pp. 109–114, 1990

Empirical Bayes Methods for Smoothing Data and for Simultaneous Estimation of Many Parameters

by Takemi Yanagimoto* and Nobuhisa Kashiwagi*

A recent successful development is found in a series of innovative, new statistical methods for smoothing data that are based on the empirical Bayes method. This paper emphasizes their practical usefulness in medical sciences and their theoretically close relationship with the problem of simultaneous estimation of parameters, depending on strata. The paper also presents two examples of analyzing epidemiological data obtained in Japan using the smoothing methods to illustrate their favorable performance.

Introduction

One of the most promising and rapidly developing branches of statistics is the use of smoothing methods that are based on the empirical Bayes approach. These methods are known in econometrics and engineering, but in medical sciences their use appears sparse in spite of their potential. The smoothing methods were developed separately from the standard statistical theory. For example, the moving average method was introduced in a heuristic way, though it is intuitively appealing. The aim of the paper is to review recent developments of smoothing methods in relation to the standard statistical method. Our emphasis will be placed on their usefulness and the need for further research on extending the methods so as to be useful in analyzing the epidemiological data.

The smoothing problem is regarded as the simultaneous estimation in a model with many strata under the assumptions that the strata are linearly ordered and the neighboring strata have density functions close to each other. This view permits us to formulate the model by describing the smoothness in terms of the prior distribution on the hyperpopulation and to embed the smoothing methods in the standard theory. Then we can construct estimators and test statistics by applying the likelihood inference such as the maximum likelihood estimator and the likelihood ratio test.

We begin with the formulation of methods in a general form, followed by the explicit description of the standard methods including the Stein problem and useful smoothing methods. Our formulation is a direct extension of well-known ones,but it is not seen in the literature. Historical notes and relations with other procedures are added. The review of smoothing methods extend to more advanced ones. Finally, our experiences in analyzing epidemiological data sets in terms of the smoothing methods are given.

*Institute of Statistical Mathematics. 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106, Japan.

Address reprint requests to T. Yanagimoto, Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106, Japan.

Methods in a General Form

Consider a model with K strata having the density (probability) function of the kth stratum, $p(x;\ \theta,\ \mu_k)$, $k = 1, \ldots, K$ where the parameter μ_k depends on the stratum and θ is common through the stratum. Let x_{ki}, $i = 1, \ldots n_k$ be a sample of size n_k from the kth stratum. Write $\mu = (\mu_1, \ldots \mu_k)$, and $x_k = (x_{k1}, \ldots x_{knk})'$. Then our problems will be the following: *a*) estimate the parameter μ, *b*) estimate the parameter θ, and *c*) test for the null hypothesis $\mu \in M_0$.

Keep in mind that our interest is placed on all the parameters in a model. We assume μ is an outcome from a hyperpopulation having the density function $g(\mu;\delta)$, $\delta \in D$, which is a prior distribution in the Bayesian context. The parameter space D has a limiting point δ_0 such that $g(\mu;\delta)$ tends to a degenerated measure; write it $g(\mu;\delta_0)$ for convenience. The null hypothesis M_0 in the test problem above will be expressed as $\delta = \delta_0$.

The overall likelihood is written as

$$L(x;\ \mu,\ \theta,\ \delta) = \{\sum_{k=1}^{K} \Pi\ p(x_{ki};\ \theta,\ \mu_k)\} g(\mu;\ \delta),$$

with $x = (x_1', \ldots, x_K')'$. Integrating the overall likelihood with respect to μ, we obtain the marginal likelihood,

$$ML(x;\ \theta,\ \delta) = \int_M L(x;\ \theta,\ \mu,\ \delta)\ d\mu$$

with M being the support of $g(\mu;\delta)$. Then our procedures are constructed as follows: *a*) estimate $\hat{\theta}$, and δ by maximizing the marginal likelihood, and *b*) estimate $\hat{\mu}$ by maximizing the (profile) overall likelihood $L(x; \mu, \hat{\theta}, \hat{\delta})$. The rejection region of the test for $\mu \in M_0$ with the level α is $T = 2 \log\{ML(x; \hat{\theta}, \hat{\delta})/ML(x; \tilde{\theta})\} > c_\alpha$, where $\tilde{\theta}$ maximizes $ML(x; \theta, \delta_0)$.

Some extensions look straightforward. The difficulties could arise in calculating the marginal likelihood, in numerical maximization of the likelihood, and also in obtaining the critical value c_α. The use of the conjugate prior distribution, if acceptable, sharply reduces computational load.

Applicable Models

Selecting density functions $p(x; \theta, \mu)$ and $g(\mu; \delta)$ suitably, we can give a variety of methods.

Example 1 (Stein Problem)

Let x_k be a sample of size 1 from a normal population $N(\mu_k, 1)$ (*1*). Suppose μ is a sample vector of size K from a normal hyperpopulation $N(0,\delta)$. In this example the common parameter θ does not appear, and the value δ_0 is 0. Then it follows that the estimate $\hat{\mu}_k = [\|x\|^2 - K]^+ x_k/\|x\|^2$ with $[z]^+ = max(z,0)$. The test statistic T takes the value 0 for $\|x\|^2 < K$ and $\|x\|^2 - K \log(\|x\|^2/K) - K$ otherwise. Therefore the rejection region of the test for $\mu_1 = \ldots = \mu_K = 0$ with a standard level α say 0.05, is $\|x\|^2 > x^2_{K;(1-\alpha)}$.

Example 2 (One-Way Design)

Let x_k be a sample vector of size n from a normal population $N(\mu_k, \sigma^2)$. Suppose μ is a sample vector of size K from a normal hyperpopulation $N(\tau,\nu)$. Then σ^2 and (τ,ν) correspond to θ and δ, respectively. Some algebras yield $\hat{\mu}_k = \bar{x} + [R-1]^+(\bar{x}_k - \bar{x})/R$ where $\bar{x}$ and $\bar{x}_k$ are sample means of x and x_k, and $R = S_b^2/S_w^2$ with S_b^2 and S_w^2 being the strata and within variances. The rejection region of the test for the homogeneity of μ_k's with a standard level is expressed as $R > F_{K-1,(n-1)K;1-\alpha'}$, which is equivalent with the conventional F test. The estimator $\hat{\sigma}^2$ is given by S_w^2 if $S_w^2 < S_b^2$ and $\{(K-1)S_b + (n-1)KS_w\}/(nK-1)$ otherwise.

The two simple examples just discussed show that the obtained estimators and tests are appealing. The derivation of methods based on other models is easily done in a parallel way, especially when the conjugate prior distribution can be assumed. However more useful methods pertain to smoothing data. We can find a series of attractive, useful methods for smoothing data, and our attention will later focus on the smoothing problem.

Example 3 (Smoothing Based on Differences of the Second Order)

In the standard smoothing problem the strata are linearly ordered in k. Let x_k be a sample of size 1 from a normal population $N(\mu_k,\sigma^2)$. To describe our confidence of gradual change of μ_k, we assume μ is an outcome from a multivariate normal hyperpopulation $N(ae_1 + \beta e_2, \delta D^-)$, where e_1 and e_2 are the normalized orthogonal vectors from $(1,\ldots,1)'$ and $(1,2,\ldots,n)'$, and D^- is the Moore-Penrose g-inverse matrix of D such that $x'Dx = \Sigma (x_{k+2} - 2x_{k+1} + x_k)^2$. Therefore it holds that $De_1 = De_2 = 0$. The null hypothesis $M_0 = \{\mu|\mu = ae_1 + be_2\}$ is expressed as $\delta_0 = 0$, consequently, $\gamma_0 = \infty$. It follows after the partial likelihood treatment that the marginal likelihood is given by

$$\log ML(\gamma) = (K-2)\log(x'(I - (I + \gamma D)^{-1})x) - (K-2)\log\gamma + \log|I + \gamma D|.$$

with $\gamma = \sigma^2/\delta$ and I being the $K \times K$ identity matrix. Let $\hat{\gamma}$ be the estimator maximizing $ML(\gamma)$. Then the estimators are $\hat{\mu} = (I + \hat{\gamma}D)^{-1}x$, $\hat{\sigma}^2 = x'(I-(I + \hat{\gamma}D)^{-1})x/(n-2)$. The rejection region of the test for linearity of μ is given by $T = 2 \log ML(\hat{\gamma})/ML(\infty) > c_\alpha$. The critical value c_α depends on K and is given using the simulation study by Yanagimoto and Yanagimoto (*2*).

The extension of the smoothing problem based on differences of the general dth order is straightforward except for obtaining c_α. The simulation studies show that critical values for $\alpha = 0.05$ are approximated by $a(d)(K + d + 1)/K$ for $d = 1, 2, 3$ and 4, where $a(1) = 2.0$, $a(2) = 1.85$, $a(3) = 1.75$ and $a(4) = 1.7$.

Historical Reviews

As far as we know, the empirical Bayesian approach to smoothing data was started by Whittaker (*3*) and Whittaker and Robinson (*4*), where the word "graduation" was used in place of "smoothing." Shiller (*5*) posed the use of the smoothness prior distribution. These gave mathematically elegant formulations of the penalized least square method. However, in these papers the estimation of the ratio of parameter $\gamma = \sigma^2/\delta$ was not given explicitly. In the Bayesian context the prior distribution is assumed to be known, but the assumption looks too restrictive in practice. Wahba and her associates (*6,7*) developed mathematical aspects of the smoothing problem and recommended the use of the generalized cross validation. The conceptual progress of likelihood inference in the Bayesian (including empirical Bayesian) model is attributed to Good (*8*). Akaike (*9*) advocated the use of type II likelihood, that is, the marginal likelihood. He also extended the smoothing problem so as to cover the seasonal adjustment.

The empirical Bayesian formulation described here is associated with various other statistical methods. Henderson (*10*) discussed the estimation problem of the component effect in random effect modeling. The procedure previously described provides an explicit one. A formal application of the EM algorithm (*11*) yields the same estimate of the parameter $\gamma = \sigma^2/\delta$. The Kalman filtering (also smoothing) is computationally efficient (*12*), though it is not easy to identify the distribution of the initial state. The practical importance of a test for homogeneity was stressed by Yanagimoto and Yanagimoto

(*2*). Morris (*13*) recommended a nomenclature, the parametric empirical Bayes method. However, it seems to the authors that a rather general term presented by Cassella (*14*) is preferable.

The smoothing model has a wide range of extensions and modifications. Later we review the seasonal adjustment model and the smoothing model in the two-dimensional space. These two models look promising in analyzing epidemiological data. Other applicable models will be found in non-Gaussian modeling. In the simultaneous estimation of many parameters as in Examples 1 and 2 the conjugate prior distribution is useful. However it is tough to develop the conjugate prior distribution in the smoothing except for the normal case, since there is no flexible, tractable, multivariate non-normal distribution (*15*). Recent researchers on non-Gaussian modeling are succeeding in innovating the analysis in this area. [For example, see West, Harrison, and Migon (*16*) and Kitagawa (*17*).] An attempt to apply the model to data for asthma attack is seen in a report by Kamakura and Yanagimoto (*18*).

Further Smoothing Methods

An advantage of the empirical Bayes smoothing method is its versatility. Actual data often has their own characteristics usable for analysis. In turn, our purpose for analyzing the data is often associated with the characteristics, for example, monthly data consisting of the incidences of diseases. (An epidemiologist may suspect a significance of the seasonal effect and hope to obtain the estimated trend.) Thus we can recommend formulating the potential seasonal effect in terms of a suitable prior distribution. Such advanced methods are still under investigation.

Example 4 (Seasonal Adjustment)

The assumptions in the general smoothing method in Example 3 are expressed as $x_k - \mu_k \sim N(0, \sigma^2)$, $\mu_{k+2} - 2\mu_{k+1} + \mu_k \sim N(0,\delta)$, $k = 1, \ldots, K-2$, $e_1'\mu = \alpha$ and $\alpha_2'\mu = \beta$. Consider a seasonal adjustment model of monthly data. The existence of seasonal effects means relative closeness of μ_k and μ_{k+12}. Obviously this requirement is not orthogonal to that of the smoothness of the trend, consequently the problem becomes much more complicated. An implementation of the seasonal adjustment is realized by assuming $\mu_k = T_k + S_k$, where

$$\begin{aligned} &S_k - S_{k+12} \sim N(0, \tau_1), && k = 1, \ldots, K\text{-}12 \\ &S_k + \ldots + S_{k+11} \sim N(0, \tau_2), && k = 1, \ldots, K\text{-}11 \\ &S_k \sim \eta_k && k = 1, \ldots, 11, \end{aligned}$$

and the requirements to T_k are the same as those in μ_k shown in Example 3. Note that we add 13 hyperparameters to the previous model. Since all the distributions appearing in this model are normal, there is no need for numerical integration for calculating the marginal likelihood. Numerical optimization is, however, still elaborate. This model was originally developed by Akaike (*9*).

The seasonal adjustment method is widely employed in econometrics and is known as a typical problem having a difficulty in identifiability. Various methods such as *X*-11 have been proposed. We again emphasize that the above approach is based on clear analytical assumptions and procedures for inference of parameters contained in the model. These advantages are mostly desirable in natural sciences.

Example 5 (Smoothing of Spatial Data)

In this example we let μ_{kh} be the variate at the (k, h)th site of a two-dimensional rectangular lattice. Whittle (*19*) proposed a simultaneous autoregressive model:

$$\mu_{kh} = \sum_{(i,j)\neq(k,h)} a_{khij}\mu_{ij} + \epsilon_{kh} \qquad \begin{aligned} k &= 1, \ldots, K, \\ h &= 1, \ldots, H, \end{aligned}$$

and applied a model of the form $x_{kh} = \mu_{kh} + \eta_{kh}$ to data for the yield of oranges obtained from uniformity trials. Here, ϵ_{kh} and η_{kh} denote a white noise, respectively. Besag (*20*) gave an errors-in-variables formulation of a conditional autoregressive model:

$$x_{kh} = \mu_{kh} + \eta_{kh}$$

$$E(\mu_{kh} \mid \text{all other values}) = \sum_{(i,j)\neq(k,h)} a_{khij}\mu_{ij}$$

$$\text{var}(\mu_{kh} \mid \text{all other values}) = \sigma_{kh}^2$$

$$\text{for any } k = 1,\ldots,K, h = 1,\ldots,H,$$

and applied it to data for the yield of wheat. The unknown parameters in both models are estimated using the maximum likelihood method. Kashiwagi (*21*) gave an empirical Bayesian formulation of a smoothing method for spatial data; he pointed out that the likelihood function, defined in both the simultaneous and conditional autoregressive models, is equivalent to the marginal likelihood function in the empirical Bayes method. In the context of the smoothing spline, Wahba (*22*) studied the use of thin plate splines for smoothing noisy multidimensional data.

It seems that this method is applicable to analyzing meshed geographical data for mobility and mortality. It enables us to give all the smoothed estimates of μ_{kh} using our knowledge of gradual changes. Descriptive methods such as the grid square method (*23*) are attributable to skilled subjective judgments.

Applications

Two examples of applying the smoothing methods to actual data obtained in Japan follow.

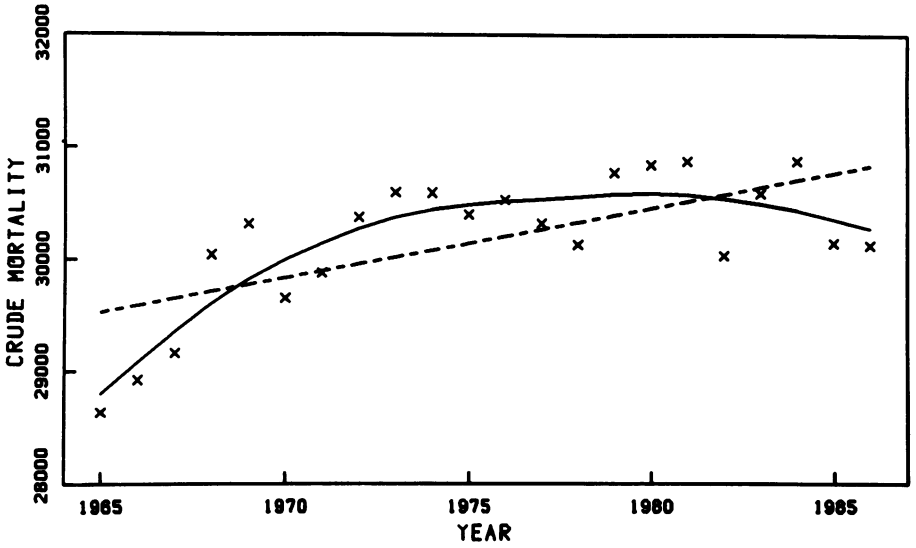

FIGURE 1. Smoothing data for annual mortality of stomach cancer in males by the empirical Bayes method (solid line) and by the simple linear regression (dotted line). $T = 11.37$.

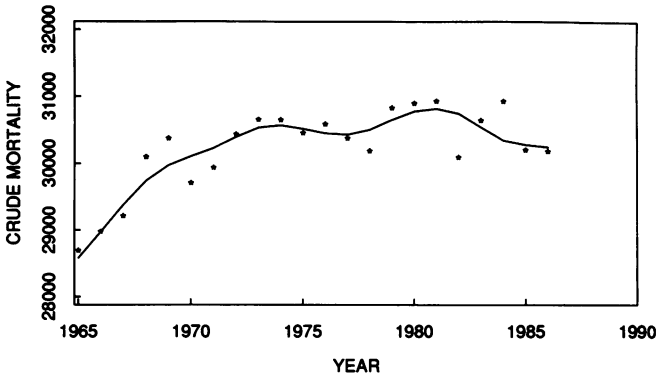

FIGURE 2. Smoothing the data as in Figure 1 by the software S.

Cancer Mortality in Japan

Stomach cancer is still the largest cause of cancer death. We analyzed yearly data cited from Japanese vital statistics for the crude number of cancer death for males during the time period between 1965 and to 1986. Figure 1 shows the result in the case of stomach cancer in males. We observe that even in the crude number base, the manual mortality has been decreasing in recent years, though it is widely accepted that the adjusted mortality is decreasing. The fitness of the simple linear regression is apparently bad. This is supported by the fact that the (marginal) likelihood ratio test statistic T takes 11.34, which is much greater than $1.85 \cdot 25/22$. To compare it with an existing method, the same data are also analyzed using the familiar statistical software, S, which is given in Figure 2. The general trends are similar, but the estimated line in Figure 2 looks overfitted; ours appears to be more appealing. A clearer difference between the two analyses is the fact that ours is closely related with the simple linear regression. The simple linear regression is powerful and often our primary choice.

We also analyzed cancer mortality data of other sites. The annual mortality of lung and pancreatic cancers of males appears to be increasing exponentially rather than linearly. Therefore, we assumed $x_k \sim LN(\mu_k,\sigma^2)$, that is, $\log x_k \sim LN(\mu_k,\sigma^2)$. Our analysis shows that the estimated lines are close to the estimated exponential regression curve. The estimated trend in lung cancer is exponential at the earlier stage of the period in the study, and it is going down from the exponential curve. On the other hand, pancreatic cancer shows better agreement with the exponential curve. However, the tests for the goodness of fit are still highly significant. The case of lung cancer is given in Figure 3.

SMON Patient Incidence

According to leading Japanese epidemiologists, subacute myelo-optico neuropathy (SMON) is a tragic

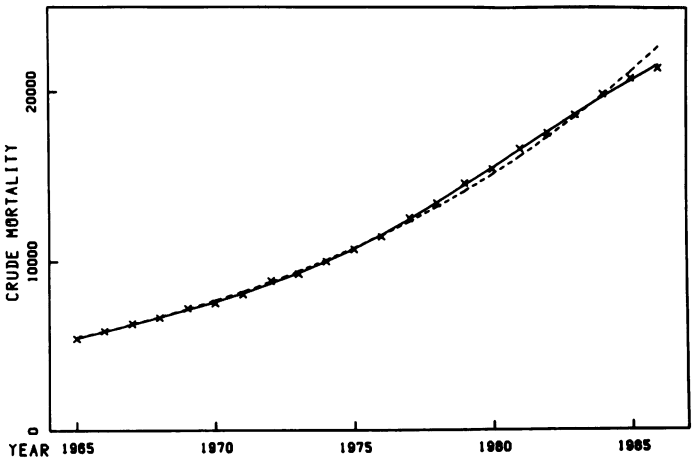

Figure 3. Smoothing data for annual mortality of lung cancer in males by the empirical Bayes method (solid line) and by the simple linear regression (dotted line), both after logarithmic transformation. $T = 18.03$.

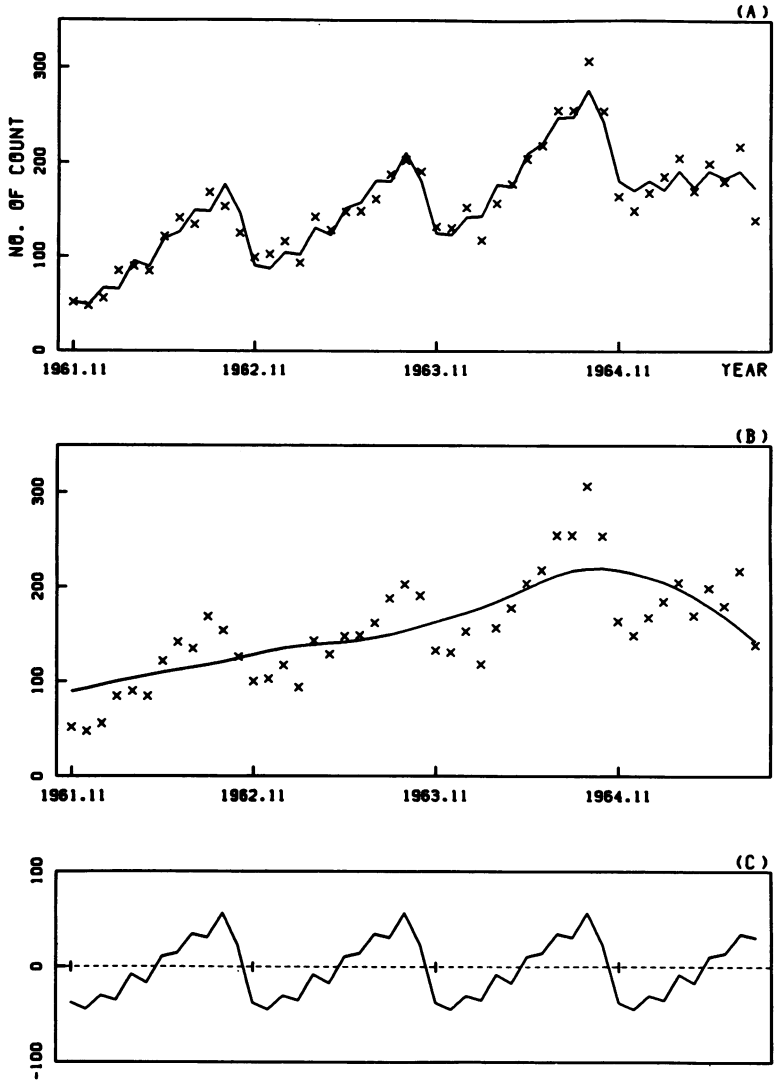

FIGURE 4. Fitting the seasonal adjustment model to data for the monthly incidence of SMON cases (*A*) with estimated general trend (*B*) and estimated seasonal factor (*C*).

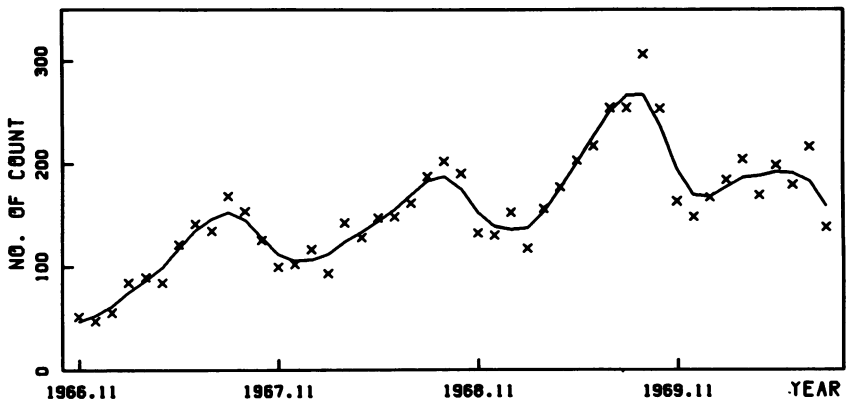

FIGURE 5. Fitting the smoothing model to the data as with Figure 4.

large-scale side effect of the drug, clioquinol. At that time when the etiology of SMON was under study, it was suspected that a relatively high incidence of SMON cases occurred in the summer. To illustrate the usefulness of the seasonal adjustment method, we analyzed the data for the monthly incidence of SMON cases cited from Table 7.1 in the Research Report (24) between November 1966 to August 1970. The estimated line with the estimated trend and seasonal effects is given in Figure 4. The smoothing model disregarding the seasonal effects is also fitted and is given in Figure 5. Both the estimated lines appear to be acceptable. More precisely, very short-term fluctuations are observed in the seasonal adjustment method. On the other hand, the upper and lower peaks cannot be interpreted well by the smoothing method. The likelihood ratio test statistic takes 50.32. Since the difference of numbers of a parameters in the models is 13, the test for the existence of seasonal effect is obviously highly significant, though we do not have explicit results on the critical value. The estimated seasonal effect shows the gradual increase of SMON from winter to summer and the highest peak seen in September, followed by a sharp decrease.

The assumption of the Poisson distribution may be more familiar than that of the normal distribution. In this case we must apply the non-Gaussian theory, and its actual implementation, including the use of computer programs, requires further investigation.

The authors thank C. Kitagawa for his guidance in the non-Gaussian approach. They also extend thanks to N. Nakajima and H. Matsuno for their help in preparing Figures 1, 2, and 3.

REFERENCES

1. Stein, C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proceedings of the 3rd Berkeley Symposium 1. University of California Press, Berkeley, CA, 1956, pp. 197–206.
2. Yanagimoto, T., and Yanagimoto, M. The use of the marginal likelihood for a diagnostic test for the goodness of fit of the simple regression model. Technometrics 29: 95–101 (1987).
3. Whittaker, E. On the new method of graduation. Proc. Edinburgh Math. Soc. 41: 62–75 (1923).
4. Whittaker, E., and Robinson, G. The Calculus of Observations. Blackie & Son, London, 1924.
5. Shiller, R. J. A distributed lag estimator derived from smoothness priors. Econometrica 41: 775–788 (1973).
6. Kimeldorf, G. S., and Wahba, G. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines, Ann. Math. Statist. 41: 495–502 (1970).
7. Golub, G. H., Heath, M., and Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21: 215–223 (1979).
8. Good, I. J. The Estimation of Probability: An Essay on Modern Bayesian Methods. The MIT Press, Cambridge, MA, 1965.
9. Akaike, H. Seasonal adjustment by a Bayesian modeling. J. Time Series Anal. 1: 1–13 (1980).
10. Henderson, C. R. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32: 69–83 (1976).
11. Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B39: 1–38 (1977).
12. Kalman, R. E. A new approach to linear filtering and prediction. ASME J. Basic Engineer. 82D: 33–45 (1960).
13. Morris, C. N. Parametric empirical Bayes inference: theory and applications. J. Am. Statist. Assoc. 78: 47–65 (1983).
14. Casella, G. An introduction to empirical Bayes data analysis. Am. Stat. 39: 83–87 (1985).
15. Yanagimoto, T. Dependence ordering in statistical models and other notions. In: Topics in Statistical Dependence (H. W. Block, A. R. Sampson, and T. H. Savitis, Eds.), in press.
16. West, M., Harrison, P. J., and Migon. H. S. Dynamic generalized linear model and Bayesian forecasting. J. Am. Stat. Assoc. 80: 73–97 (1985).
17. Kitagawa, G. Non-Gaussian state-space modeling of nonstationary time series. J. Am. Stat. Math. 82: 1032–1063 (1987).
18. Kamakura, T., and Yanagimoto, T. Point process models in asthma attacks for environmental factors. Environ. Health Perspect. 63: 203–210 (1985).
19. Whittle, P. On stationary processes in the plane. Biometrika 41: 434–449 (1954).
20. Besag, J. E. Errors-in-variables estimation for Gaussian lattice schemes. J. R. Stat. Soc. B39: 73–78 (1977).
21. Kashiwagi, N. Estimation of fertilities in field experiments [in Japanese]. Proc. Inst. Stat. Math. 30: 1–10 (1982).
22. Wahba, G. Convergence rates of "thin plate" smoothing splines. In: Smoothing Techniques for Curve Estimation (T. Gasser and M. Rosenblatt, Eds.), Springer-Verlag, New York, 1979, pp. 233–245.
23. Ohkubo, T. Study of cancer mortality by grid square method. Environ. Health Perspect. 32: 75–81 (1979).
24. The Subacute Myleo-Optical Neuropathy Research Commission, Epidemiology Group. Research Report FY1971 No. 8, Soft Science Publications, Tokyo, Japan, 1972.